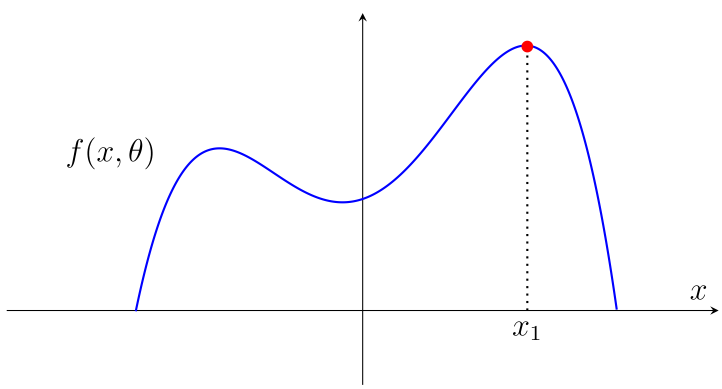
LABORATOR #11

EX#1 (Aproximare de parametri folosind Estimatorul de Verosimilitate Maximă) Presupunem că avem $x_1, x_2, \dots, x_N$ simulări independente generate dintr-o distribuție V_θ cu o densitate $f(x, \theta)$, ce depinde de un parametru necunoscut $\theta \in \mathbb{R}$. Dorim să aproximăm θ plecând de la setul de date $\{x_1, x_2, \dots, x_N\}$.

Exemplu: Fie $x_1, x_2, \dots, x_N$ simulări independente din distribuția $\mathcal{N}(\theta, 1)$. Cum determinăm θ folosind setul de date (dorim o metodă care să funcționeze chiar dacă θ nu este media sau varianța distribuției, pe care știm să le aproximăm prin Legea Numerelor Mari)?

Ne întoarcem la cadrul general:

- Dacă am avea doar o singură simulare x_1, atunci θ cel mai verosimil este acela pentru care $f(x_1, \theta)$ este maxim.



- Dacă avem două simulări independente x_1, x_2, acest lucru este echivalent cu o singură simulare din distribuția $V_\theta \times V_\theta$, a cărei densitate se poate calcula astfel: fie X_1, X_2 independente distribuite conform V_θ. Atunci, pentru orice $[a_1, b_1]$ și $[a_2, b_2]$ intervale reale,

$$\begin{aligned}\mathbb{P}\Big((X_1, X_2) \in [a_1, b_1] \times [a_2, b_2]\Big) &= \mathbb{P}(X_1 \in [a_1, b_1])\, \mathbb{P}(X_2 \in [a_2, b_2]) \\ &= \int_{a_1}^{b_1} f(x, \theta) \mathrm{d}x \int_{a_2}^{b_2} f(y, \theta) \mathrm{d}y \\ &= \int_{[a_1,b_1]\times[a_2,b_2]} f(x, \theta) f(y, \theta) \mathrm{d}x \mathrm{d}y.\end{aligned}$$

  Prin urmare, densitatea distribuției $V_\theta \times V_\theta$ este:

$$f_2((x, y), \theta) \coloneqq f(x, \theta) f(y, \theta)$$

  Prin urmare, date fiind x_1 și x_2, cea mai verosimilă valoare pentru θ este aceea care maximizează funcția de densitate $f_2((x_1, x_2), \theta)$.

- Dacă avem N simulări, cea mai verosimilă valoare pentru θ este:

$$\theta^* = \underset{\theta}{\operatorname{argmax}}\, f_N((x_1, x_2, \dots, x_N), \theta),$$

unde $f_N((x_1, x_2, \dots, x_N), \theta) \coloneqq f(x_1, \theta) f(x_2, \theta) \cdots f(x_N, \theta)$.

Notă: Funcția $\mathcal{L}_N(\theta) \coloneqq f_N((x_1, x_2, \dots x_N), \theta)$ se numește funcția de verosimilitate a parametrului θ în raport cu setul de date $\{x_1, x_2, \dots x_N\}$. De multe ori, în practică este preferată maximizarea funcției:

$$\log \mathcal{L}_N(\theta) \coloneqq \sum_{i=1}^{N} \log f(x_i, \theta),$$

iar $\operatorname{argmax}_\theta \mathcal{L}_N(\theta)$ și $\operatorname{argmax}_\theta \log \mathcal{L}_N(\theta)$ coincid deoarece funcția log este strict crescătoare.

Exemplu: Fie $x_1, x_2, \dots, x_N$ simulări independente din distribuția $\mathcal{N}(\theta, 1)$. În acest caz,

$$f(x, \theta) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-\theta)^2}{2}},$$

iar logaritmul funcției de verosimilitate are forma:

$$\log \mathcal{L}_N(\theta) = N \log \left(\frac{1}{\sqrt{2\pi}} \right) - \frac{1}{2} \sum_{i=1}^{N} (x_i - \theta)^2.$$

Ignorând constanta aditivă, trebuie să maximizăm funcția:

$$\theta \to \left(-N\theta^2 + 2\theta \sum_{i=1}^{N} x_i - \sum_{i=1}^{N} x_i^2 \right).$$

Din proprietățile fundamentale ale funcției de gradul al doilea, obținem:

$$\theta^* = \frac{\sum_{i=1}^{N} x_i}{N} = \bar{x}_N.$$

#1 Pentru $N = 10000$, simulați numerele reale $y_1, y_2, \dots, y_N$ independente și distribuite $\mathcal{N}(0, 1)$ și, independent de acestea, numerele $z_1, z_2, \dots, z_N$ independente și care iau valorile ± 1 cu probabilitate egală. Pentru un număr $\theta \in (0, 1)$, afișați histograma valorilor $x_i \coloneqq \theta y_i + \sqrt{1 - \theta^2}\, z_i$, $i = \overline{1, N}$.

#2 Determinați teoretic funcția de densitate a distribuției V_θ a numerelor $(x_i)_{i=1}^N$ generate la subpunctul anterior și afișați graficul acesteia peste histogramă.

Hint: Fie $h(x) \coloneqq \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$ densitatea distribuției $\mathcal{N}(0, 1)$. Fie $X = \theta Y + \sqrt{1 - \theta^2}\, Z$, unde $Y \sim \mathcal{N}(0, 1)$ și $Z \sim \left[2 \operatorname{Bernoulli}\left(\frac{1}{2}\right) - 1\right]$ sunt independente. Pentru $a < b$ arbitrare, calculăm $\mathbb{P}(X \in [a, b])$ astfel:

$$\begin{aligned}
\mathbb{P}\left(\theta Y + \sqrt{1-\theta^2}\, Z \in [a, b]\right) &= \mathbb{P}\left(\theta Y + \sqrt{1-\theta^2} Z \in [a, b] \cap Z = 1\right) \\
&\quad + \mathbb{P}\left(\theta Y + \sqrt{1-\theta^2} Z \in [a, b] \cap Z = -1\right) \\
&= \mathbb{P}\left(\theta Y \in \left[a - \sqrt{1-\theta^2}, b - \sqrt{1-\theta^2}\right] \cap Z = 1\right) \\
&\quad + \mathbb{P}\left(\theta Y \in \left[a + \sqrt{1-\theta^2}, b + \sqrt{1-\theta^2}\right] \cap Z = -1\right).
\end{aligned}$$

Din independența variabilelor Y și Z obținem:

$$\begin{aligned}
\mathbb{P}\left(\theta Y + \sqrt{1-\theta^2}\, Z \in [a,b]\right) &= \mathbb{P}\left(\theta Y \in \left[a - \sqrt{1-\theta^2}, b - \sqrt{1-\theta^2}\right]\right) \mathbb{P}(Z=1) \\
&\quad + \mathbb{P}\left(\theta Y \in \left[a + \sqrt{1-\theta^2}, b + \sqrt{1-\theta^2}\right]\right) \mathbb{P}(Z=-1) \\
&= \frac{1}{2}\mathbb{P}\left(Y \in \left[\frac{a - \sqrt{1-\theta^2}}{\theta}, \frac{b - \sqrt{1-\theta^2}}{\theta}\right]\right) \\
&\quad + \frac{1}{2}\mathbb{P}\left(Y \in \left[\frac{a + \sqrt{1-\theta^2}}{\theta}, \frac{b + \sqrt{1-\theta^2}}{\theta}\right]\right) \\
&= \frac{1}{2}\int_{\frac{a-\sqrt{1-\theta^2}}{\theta}}^{\frac{b-\sqrt{1-\theta^2}}{\theta}} h(t)\mathrm{d}t + \frac{1}{2}\int_{\frac{a+\sqrt{1-\theta^2}}{\theta}}^{\frac{b+\sqrt{1-\theta^2}}{\theta}} h(t)\mathrm{d}t.
\end{aligned}$$

Efectuând schimbarea de variabilă $t = \frac{s \pm \sqrt{1-\theta^2}}{\theta}$, obținem:

$$\mathbb{P}\left(\theta Y + \sqrt{1-\theta^2}\, Z \in [a,b]\right) = \int_a^b \frac{1}{2\theta}\left[h\left(\frac{s - \sqrt{1-\theta^2}}{\theta}\right) + h\left(\frac{s + \sqrt{1-\theta^2}}{\theta}\right)\right] \mathrm{d}s.$$

Prin urmare, densitatea distribuției V_θ este:

$$f(x,\theta) = \frac{1}{2\theta}\left[h\left(\frac{x - \sqrt{1-\theta^2}}{\theta}\right) + h\left(\frac{x + \sqrt{1-\theta^2}}{\theta}\right)\right].$$

#3 Recuperați parametrul θ folosind exclusiv setul de date generat $\{x_1, x_2, \ldots, x_N\}$ folosind Estimatorul de Verosimilitate Maximă. Afișați graficul logaritmului funcției de verosimilitate, marcând printr-o linie punctată verticală valoarea θ^* în care acesta își atinge maximul.
Hint: Folosiți funcția de optimizare `minimize_scalar` din biblioteca `scipy.optimize`.